

2017 Machine Learning with R

Association Rules

강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

목차

I

연관규칙분석:A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

한번쯤은 봤을 법한...

R을 이용한 데이터마이닝
무료배송

박창이, 김용대, 김진석, 송종우, 최호식 지음 | 교우사 | 2011년 07월 30일 출간

크게보기

이 책의 영업점 진열 위치
광화문 | 강남 | 잠실 | 분당 | 목동

다른영업점 보기
영업점 전체재고

0
0
0
0

정가 : 26,000원
판매가 : **26,000원** [0%+0원 할인] 정가제 Free
청구할인가 : **23,400원** 최대 10%할인 현대카드 M포인트 결제할인 안내
통합포인트 : 260원 [1%적립] 안내 **OH! POINT** 0.3% 추가적립 안내

2014년 전국연합학력평가시행 **고등학교 전학년 6월 모의고사 완벽대비**

· 배송비 : 무료 배송비 안내

· 도착예정일 : 지금 주문하면 **내일(13일, 화) 도착 예정** > 도착예정일 안내
서울 종로구 종로1가 교보생명빌딩 지역변경

서울,수도권	부산	대구,창원
24시까지 주문하면 내일(13일, 화) 도착	24시까지 주문하면 5월14일(수)이내 도착	24시까지 주문하면 5월14일(수)이내 도착

· 바로드림 : 인터넷 주문 **1시간 후 선택 영업점**에서 직접수령 안내 **바로드림주문하기**

신용카드 할인

KB
LOTTECARD
ingood!Card
> 더보기

포인트 적립

pointree
B
TOP
citi
M
> 더보기

주문수량 1

최저가보상
해외배송가능
프리미엄배송

장바구니 담기
바로구매
보관함에 담기

한번쯤은 봤을 법한...

R을 이용한 데이터마이닝

무료배송

박창이, 김용대, 김진석, 송종우, 최호식 지음 | 교우사 | 2011년 07월 30일 출간



정 가 : 26,000원

판 매 가 : **26,000원** [0%+0원 할인] 정가제 Free

청구할인가 : **23,400원** 최대 10%할인 현대카드 M포인트 결제할인 안내

통합포인트 : 260원 [1%적립] 안내 OH! POINT 0.3% 추가적립 안내

이 책과 함께 구매한 도서

▲ 위로

☐ 전체선택

장바구니에 담기

보관함에 담기



데이터마이닝 방법론(박창이 R로 배우는 데이터 분석
이터 분석을 ...
25,000원
[0%+1%P]



데이터마이닝 입문(SAS 데이터마이닝(비즈니스 인
터프...
28,800원
[10%+10%P]



데이터마이닝 입문(SAS 데이터마이닝(비즈니스 인
터프...
20,000원
[0%+1%P]



데이터마이닝(SAS 데이터마이닝(비즈니스 인
터프...
33,000원
[0%+1%P]



데이터마이닝(R SAS
MS-SQL을...
30,000원
[0%+1%P]

한번쯤은 봤을 법한...

amazon [Try Prime](#) Pilsung's Amazon.com Today's Deals Gift Cards Sell Help

Shop by Department [Electronics](#) iPad [Go](#) Hello, Pilsung [Your Account](#)

[Computers](#) [Brands](#) [Best Sellers](#) [Laptops & Tablets](#) [Desktops & Monitors](#) [Hard Drives & Storage](#) [Computer Accessories](#) [Ta](#)

[Electronics](#) > [Computers & Accessories](#) > [Tablets](#)



Apple iPad Air 2 MH182LL/A (64GB, Wi-Fi, Gold) NEWEST VERSION
by Apple
★★★★★ 739 customer reviews
| 27 answered questions

Price: **\$583.00**

In Stock.
Sold by [TechGiant](#) and Fulfilled by Amazon. Gift-wrap available.

This item does not ship to **Seoul, Korea; Republic of (South Korea)**. Please check other sellers who may ship internationally.
[Learn more](#)

Size: **64 GB**

[16 GB](#) [64 GB](#) [128 GB](#)

Frequently Bought Together





Price for all three: \$601.84

[Add all three to Cart](#)

[Add all three to Wish List](#)

[Show availability and shipping details](#)

- ☒ **This item:** Apple iPad Air 2 MH182LL/A (64GB, Wi-Fi, Gold) NEWEST VERSION **\$583.00**
- ☒ Tech Armor Apple iPad Air 2 / iPad Air (first generation) High Definition (HD) Clear Screen Protectors ... **\$8.95**
- ☒ Apple iPad Air 2 Case - MoKo Ultra Slim Lightweight Smart-shell Stand Cover Case for Apple iPad Air 2 ... **\$9.89**

일상 생활에서의 연관 규칙 분석

❖ 장바구니 분석: Market basket analysis (MBA)



Wall Mart (USA)



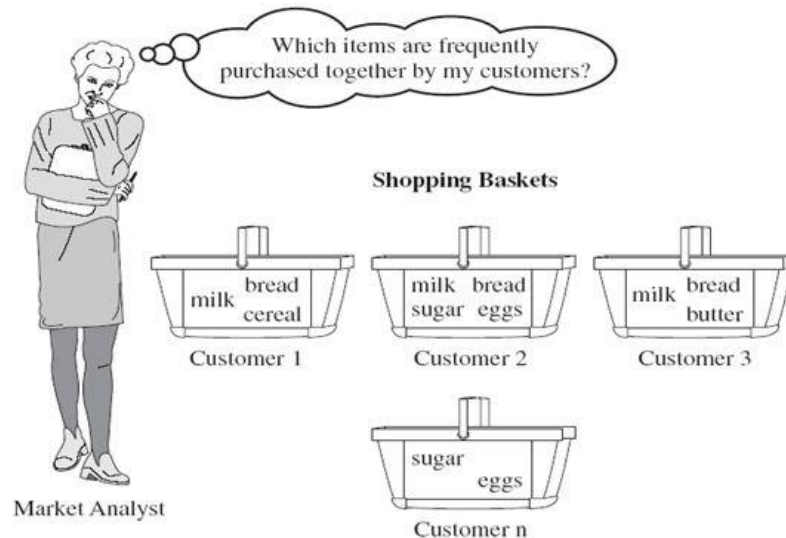
E-Mart (Korea)



연관규칙 분석

❖ 목적

- 어떤 두 아이템 집합이 **빈번히 발생**하는가를 알려주는 **일련의 규칙**들을 생성
 - ✓ Produce rules that define “what goes with what”
- 우리의 데이터에 의하면 “X 아이템을 구매하는 고객들은 Y 아이템 역시 구매할 가능성이 높다” → 콘텐츠 기반 추천 시스템에 널리 사용
- 장바구니 분석(Market Basket Analysis)으로도 널리 알려짐



연관규칙 분석

❖ 데이터 속성

- 각 레코드는 트랜잭션의 형태를 가짐
- 행렬의 형태로 표현하게 되면 대부분의 셀이 0의 값은 갖는 희소행렬(sparse matrix)이 됨

Tid	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Tid	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

연관규칙 분석: 예제

❖ 동네 작은 가게 매출 트랜잭션 데이터

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

연관규칙 분석: 용어 및 규칙 생성

❖ 용어: Terminology

- 조건절(Antecedent) – “IF” part
- 결과절(Consequent) – “THEN” part
- 아이템 집합(Item set) – 조건절 또는 결과절을 구성하는 아이템들의 집합
- 조건절 아이템 집합과 결과절 아이템 집합은 상호배반 (한 아이템이 조건절과 결과절에 모두 포함될 수 없음)

❖ 규칙 생성: Generating rules

- 매우 많은 수의 규칙이 생성 가능 (예시: 첫번째 트랜잭션)
 - ✓ 계란을 구매하는 사람들은 라면도 함께 구매한다.
 - ✓ 계란과 라면을 구매하는 사람들은 참치도 함께 구매한다.
 - ✓ 참치를 구매하는 사람들은 계란도 함께 구매한다.
 - ✓ ...

연관규칙 분석: 규칙의 효용성 측정 지표

For the rule $A \rightarrow B$

❖ 지지도: Support

$$\text{support}(A) = P(A)$$

- 빈발 아이템 집합(frequent item sets)을 판별하는데 사용

❖ 신뢰도: Confidence

$$\text{confidence}(A \rightarrow B) = \frac{P(A, B)}{P(A)}$$

- 아이템 집합 간의 연관성 강도를 측정하는데 사용

❖ 향상도: Lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

- 생성된 규칙이 실제 효용가치가 있는지를 판별하는데 사용

연관규칙 분석: 규칙의 효용성 측정 지표

Rule: $X \Rightarrow Y$

$$Support = \frac{freq(X, Y)}{N}$$

$$Confidence = \frac{freq(X, Y)}{freq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

연관규칙 분석: 규칙 생성

❖ 유용한 연관 규칙들을 어떻게 찾아낼 것인가?

- 이상적으로는 모든 생성 가능한 규칙을 만든 뒤, 각 규칙의 지지도, 신뢰도, 향상도를 측정하여 유용한 규칙들만을 찾아냄
- 아이템 수가 증가할수록 계산에 소요되는 시간이 기하급수적으로 증가함

❖ Brute-force approach

- 가능한 모든 규칙을 나열함
- 모든 규칙의 지지도와 신뢰도를 계산함
- 최소지지도와 최소신뢰도 조건을 만족하지 못하는 규칙을 제거
- Computationally prohibitive!

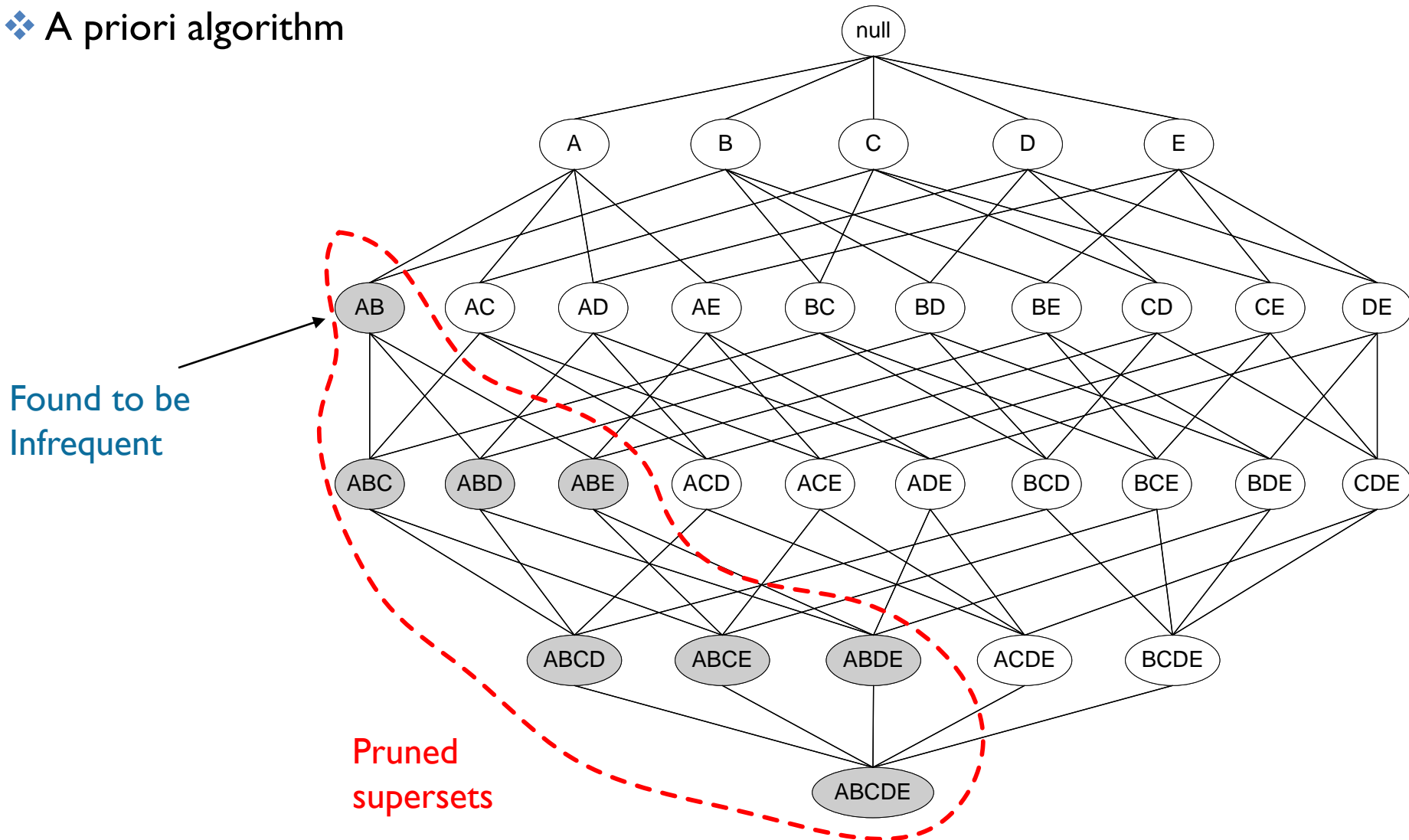
연관규칙 분석: 규칙 생성

❖ A priori algorithm

- 빈발 집합(frequent item sets)만을 고려하여 규칙 생성
- **지지도(support)**
 - ✓ 조건절에 속하는 아이템 집합이 발생할 확률
 - ✓ 아이템 집합 {계란, 라면}의 지지도는 40%
- **최소 지지도(minimum support)**
 - ✓ 유용한 규칙으로 인정받기 위해 필요한 최소 지지도
- 최소 지지를 만족하지 못하는 아이템 집합의 상위집합(superset)은 항상 최소 지지를 만족하지 않음
 - ✓ Support of an item set never exceeds the support of its subsets, which is known as **anti-monotone** property of support.

연관규칙 분석: 규칙 생성

❖ A priori algorithm



연관규칙 분석: 빈발 아이템 집합 생성

최소 지지도 조건 부여

- 최소 지지도: 2 transactions or 20%

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

연관규칙 분석: 빈발 아이템 집합 생성

2

최소 지지도 조건을 만족하는 1개짜리 아이템 집합을 생성

- $\text{Support \{noodle\}} = 8/10 = 80\%$
- $\text{Support \{egg\}} = 5/10 = 50\%$
- $\text{Support \{cola\}} = 5/10 = 50\%$
- $\text{Support \{rice\}} = 3/10 = 30\%$
- $\text{Support \{tuna\}} = 2/10 = 20\%$
- $\text{Support \{onion\}} = 1/10 = 10\%$

양파(onion)는 최소 지지도 조건을 만족하지 못했으므로 이후 분석에서 제외

연관규칙 분석: 빈발 아이템 집합 생성

3

앞 단계에서 살아남은 아이템들을 이용하여 최소 지지도 조건을 만족하는 2개짜리 아이템 집합을 생성

	noodle	egg	cola	rice	tuna
noodle		40%	40%	20%	20%
egg			30%	0%	20%
cola				0%	10%
rice					0%
tuna					

- {noodle, egg}, {noodle, cola}, {noodle, rice}, {noodle, tuna}, {egg, cola}, {egg, tuna} are frequent two-item sets.

연관규칙 분석: 빈발 아이템 집합 생성

더 이상 최소 지지도 이상을 나타내는 아이템 집합이 없을 때까지 아이템 집합의 크기를 1씩 증가시키면서 반복 수행

4

Set-size	Item 1	Item 2	Item 3	...	Item 6
1	noodle				
1	egg				
1	cola				
1	rice				
1	tuna				
2	noodle	egg			
2	noodle	cola			
2	noodle	rice			
...			

연관규칙 분석: A Priori Algorithm

❖ A priori algorithm

- Let $k=1$
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
 - ✓ Generate length $(k+1)$ candidate itemsets from length k frequent itemsets
 - ✓ Prune candidate itemsets containing subsets of length k that are infrequent
 - ✓ Count the support of each candidate by scanning the DB
 - ✓ Eliminate candidates that are infrequent, leaving only those that are frequent

연관규칙 분석: 규칙 평가

❖ 신뢰도: Confidence

- 조건절이 발생했다는 가정 하에 결과절이 발생할 조건부 확률
- E.g. “if noodle is purchased, then egg is also purchased”

$$\text{support}(\text{noodle}) = P(\text{noodle}) = \frac{8}{10}, \quad \text{support}(\text{egg}) = P(\text{egg}) = \frac{5}{10}$$

$$\text{confidence}(\text{noodle} \rightarrow \text{egg}) = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})} = \frac{4/10}{8/10} = 0.5(50\%)$$

- 비교 대상 신뢰도(benchmark confidence): 전체 트랜잭션에서 결과절이 발생할 확률 ($P(\text{egg})$, $\text{support}(\text{egg})$)
- 규칙 ($\text{noodle} \rightarrow \text{egg}$)의 신뢰도가 $P(\text{egg})$ 보다 작으면 규칙으로서의 효용 가치는 낮음

연관규칙 분석: 규칙 평가

❖ 지지도: Lift

- 신뢰도/비교 대상 신뢰도: Confidence/(benchmark confidence)

$lift(noodle \rightarrow egg)$

$$\begin{aligned}
 &= \frac{confidence(noodle \rightarrow egg)}{support(egg)} = \frac{P(noodle, egg)}{P(egg)} = \frac{P(noodle, egg)}{P(noodle) \times P(egg)} \\
 &= \frac{\frac{4}{10}}{\frac{8}{10} \times \frac{5}{10}} = 1
 \end{aligned}$$

- 신뢰도 = 1일 경우, 조건절과 결과절은 통계적으로 독립사건임을 의미함 → 규칙 사이에 유의미한 연관성이 없음
- 신뢰도 > 1일 경우 조건절과 결과절은 서로 긍정적인 연관관계를 나타냄

연관규칙 분석: 사례 결과

❖ 규칙 생성을 위한 기준 지지도 및 신뢰도 설정

- 기준 지지도: 20%
- 기준 신뢰도: 70%

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

연관규칙 분석: 요약

❖ 연관규칙분석

- 트랜잭션 데이터베이스에 존재하는 아이템 집합들 간의 연관성을 나타내는 규칙을 생성하는 분석 기법
- 다양한 분야의 추천 시스템 구축에 널리 사용됨
- 전체 규칙을 모두 생성하는 것이 비효율적이기 때문에 효율적인 빈발 집합을 찾아내는 A Priori 알고리즘을 사용
- 규칙의 효용성은 지지도, 신뢰도, 향상도의 세 가지를 이용하여 평가
- 규칙 1: $A \rightarrow B$ 와 규칙 2: $C \rightarrow D$ 에 대해 지지도, 신뢰도, 향상도가 모두 클 경우에만 규칙 1이 규칙 2보다 효과적인 규칙으로 결론지을 수 있음

목차

I

연관규칙분석:A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

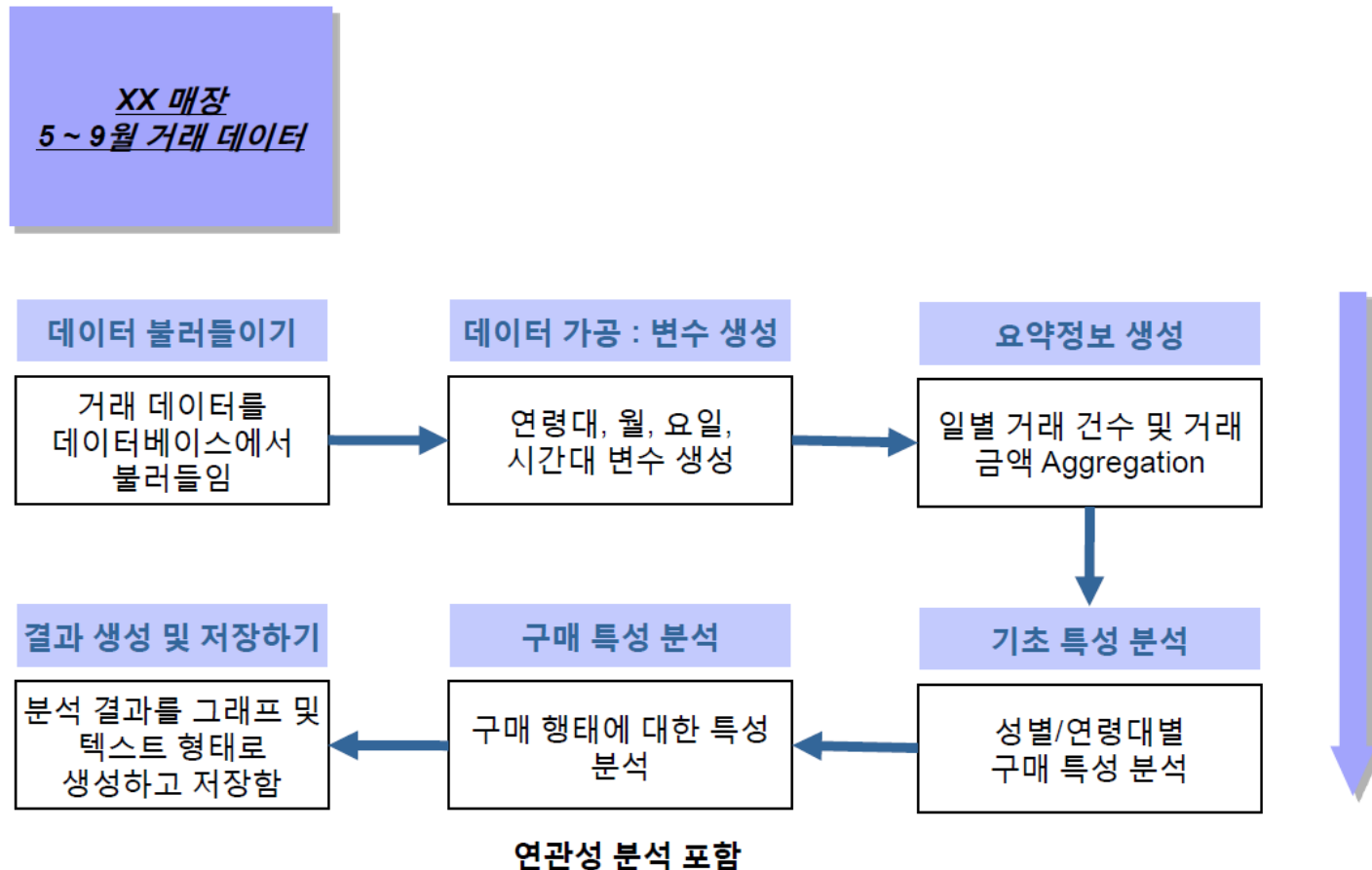
■ 데이터 구조

변수명	설명
회원번호	개인 고유 식별 번호
성별	남자 or 여자
연령대	회원의 생년월일을 기준으로 연령대 구분
발급일자	회원 가입일자
매출일자	상품 구매 일자
매출요일	상품 구매 요일
평일공휴일구분	평일 혹은 공휴일 구분 (공휴일은 토요일, 일요일, 휴일)
대분류명	상품의 대분류 구분
중분류명	상품의 중분류 구분
소분류명	상품의 소분류 구분
세분류명	상품의 세분류 구분
수량	상품 구매 개수
매출금액	상품 구매 금액

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 데이터 분석 프로세스



사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 데이터 분석 시나리오

	<u>시나리오 명</u>	<u>분석 방향</u>
시나리오 1	기초 특성 분석	<ul style="list-style-type: none"> • 상세 분석을 위한 데이터 파악 • 구매 현황 리포트 작성
시나리오 2	그룹별 구매 패턴 분석	<ul style="list-style-type: none"> • 전체, 성별/연령대별 상품 구매 패턴 파악
시나리오 3	개인별 구매 패턴 분석	<ul style="list-style-type: none"> • 개인별 상품 구매 패턴 파악
시나리오 4	연관성 분석	<ul style="list-style-type: none"> • 주로 구매하는 상품으로 장바구니를 구성함

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

▪ 구매 물품의 시각화

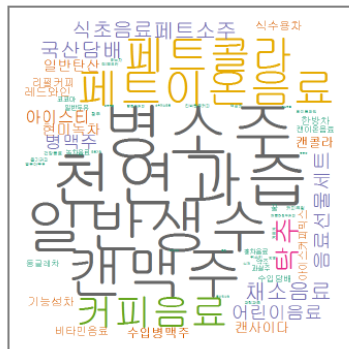


사례 1: 유통에서의 연관규칙분석

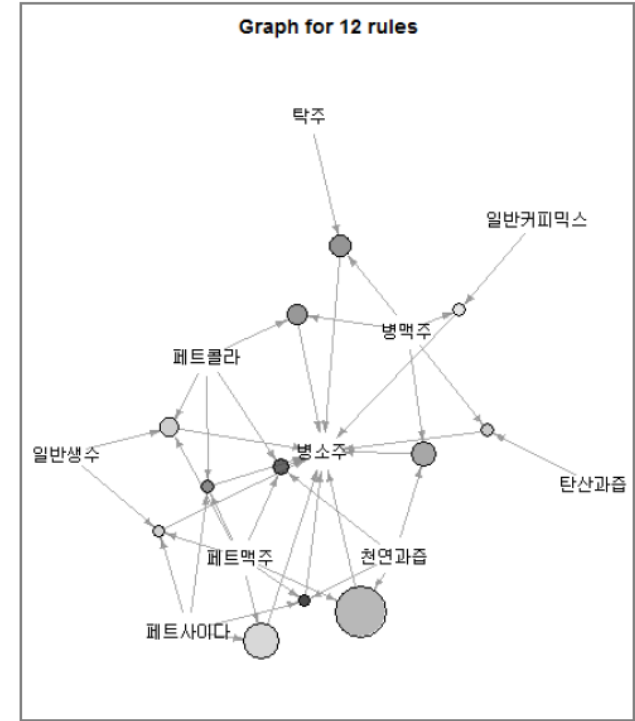
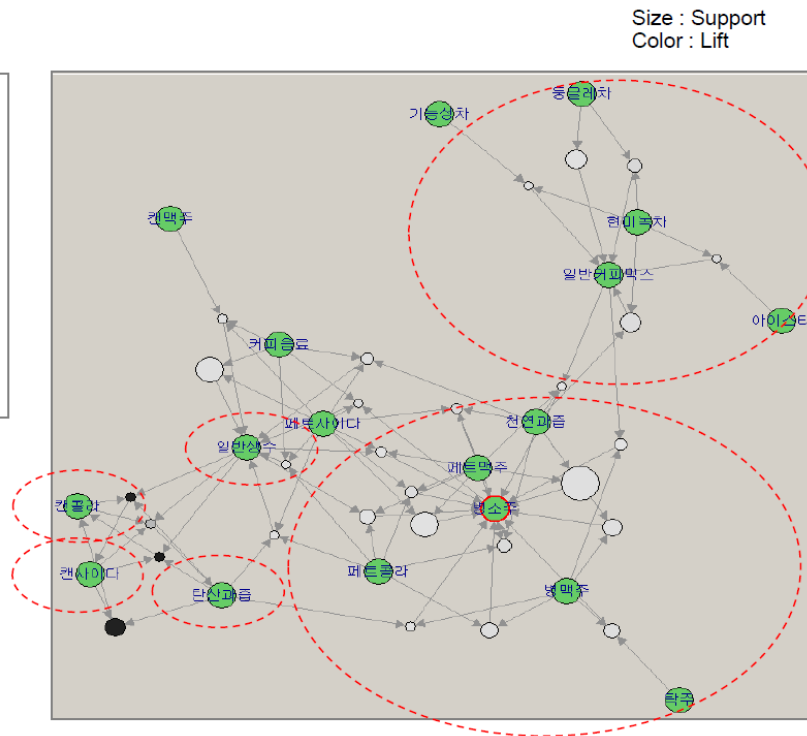
❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 연관규칙 분석을 통한 상품집합 도출

● 주류,기호



캔콜라	2
캔사이다	1
탄산과즙	1
일반커피믹스	5
일반생수	6
병소주	13



- 연구 목적: 근접무선통신기술(NFC)을 이용하여 수집한 박람회 관람객 행동 데이터를 분석하여 전시부스간 연관관계 분석 및 부스 배치 최적화
- 데이터 수집 체계



사례 2: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

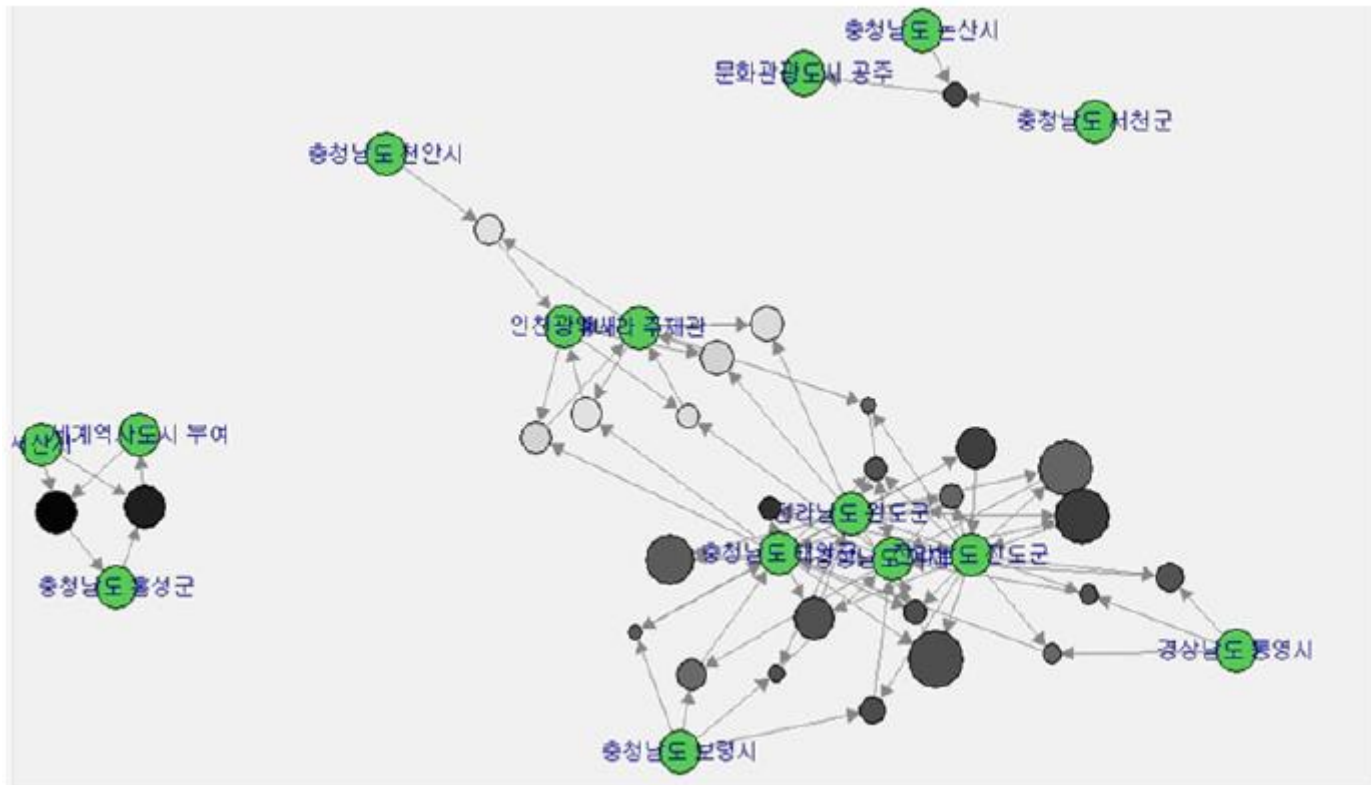
■ 연관규칙 분석 결과

	rules	support	confidence	lift
1	{충청남도 논산시, 충청남도 서천군} => {문화관광도시 공주}	0.09631728	0.8500000	4.148157
2	{서산시, 세계역사도시 부여} => {충청남도 홍성군}	0.10859301	0.9274194	5.280307
3	{서산시, 충청남도 홍성군} => {세계역사도시 부여}	0.10859301	0.8156028	4.798463
4	{전라남도 진도군, 충청남도 보령시} => {경상남도 거제시}	0.09820585	0.8888889	4.040057
5	{전라남도 완도군, 충청남도 보령시} => {경상남도 거제시}	0.09254013	0.8750000	3.976931
6	{전라남도 완도군, 충청남도 보령시} => {충청남도 태안군}	0.09065156	0.8571429	3.830018
7	{경상남도 거제시, 충청남도 보령시} => {충청남도 태안군}	0.10103872	0.8045113	3.594842
8	{경상남도 통영시, 전라남도 진도군} => {경상남도 거제시}	0.10009443	0.8688525	3.948990
9	{경상남도 통영시, 전라남도 진도군} => {충청남도 태안군}	0.09348442	0.8114754	3.625960
10	{경상남도 통영시, 전라남도 완도군} => {경상남도 거제시}	0.09348442	0.8761062	3.981959
11	{전라남도 완도군, 전라남도 진도군} => {경상남도 거제시}	0.11709160	0.8104575	3.683582
12	{경상남도 거제시, 전라남도 완도군} => {전라남도 진도군}	0.11709160	0.8157895	4.319605
13	{전라남도 진도군, 충청남도 태안군} => {전라남도 완도군}	0.10859301	0.8156028	3.980292
14	{전라남도 완도군, 충청남도 태안군} => {전라남도 진도군}	0.10859301	0.8041958	4.258217
15	{전라남도 진도군, 충청남도 태안군} => {경상남도 거제시}	0.11803588	0.8865248	4.029312
16	{내나라 주재관, 전라남도 진도군} => {경상남도 거제시}	0.09159585	0.8151261	3.704800
17	{전라남도 완도군, 충청남도 태안군} => {경상남도 거제시}	0.11331445	0.8391608	3.814040
18	{내나라 주재관, 전라남도 완도군} => {인천광역시}	0.10387158	0.8396947	1.967338
19	{인천광역시, 전라남도 완도군} => {내나라 주재관}	0.10387158	0.8661417	2.084646
20	{내나라 주재관, 충청남도 천안시} => {인천광역시}	0.10103872	0.8045113	1.884906
21	{경상남도 거제시, 인천광역시} => {내나라 주재관}	0.09631728	0.8225806	1.979802
22	{내나라 주재관, 충청남도 태안군} => {인천광역시}	0.10198300	0.8059701	1.888324
23	{인천광역시, 충청남도 태안군} => {내나라 주재관}	0.10198300	0.8571429	2.062987
24	{경상남도 거제시, 전라남도 완도군, 전라남도 진도군} => {충청남도 태안군}	0.09631728	0.8225806	3.675582
25	{전라남도 완도군, 전라남도 진도군, 충청남도 태안군} => {경상남도 거제시}	0.09631728	0.8869565	4.031274
26	{경상남도 거제시, 전라남도 진도군, 충청남도 태안군} => {전라남도 완도군}	0.09631728	0.8160000	3.982230
27	{경상남도 거제시, 전라남도 완도군, 충청남도 태안군} => {전라남도 진도군}	0.09631728	0.8500000	4.500750

사례 2: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

■ 연관규칙 분석 결과

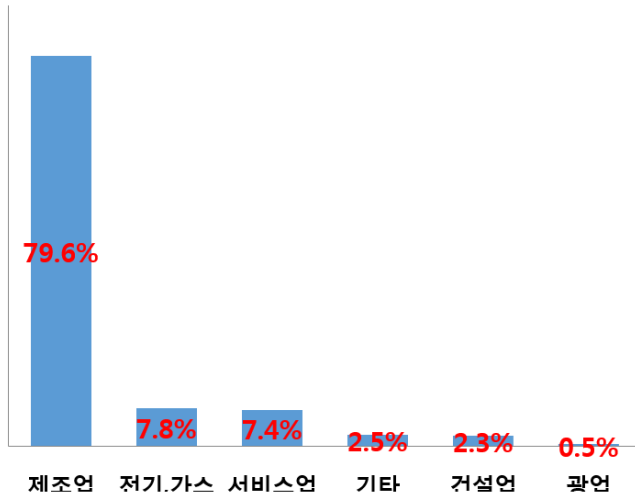


사례 3: 기업체 교육프로그램 추천

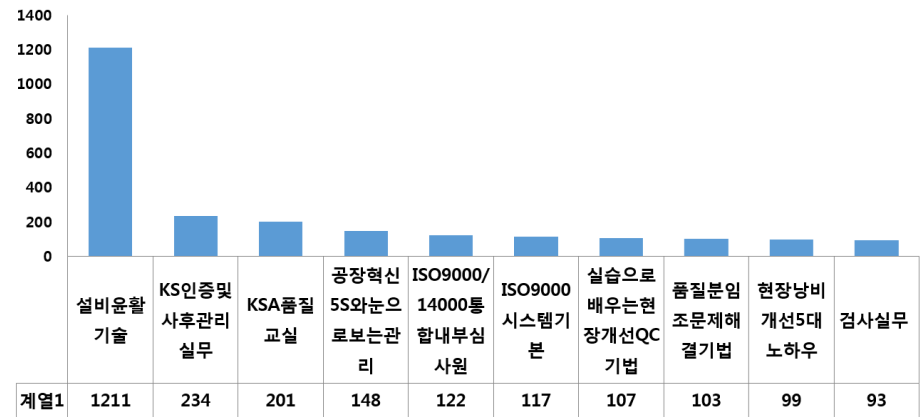
❖ 국내 K사의 기업체 교육 프로그램 운영 현황

- 교육 대상 기업의 주 업종은 제조업으로써 전체 교육 이수 기업의 79.6%에 해당함
- 대부분의 기업이 1~5개의 교육을 이수하고 있음

교육기관	기간	참여기업(수)	교육프로그램(수)	참가횟수
K사	2012 년 1월~8월	3,604	263	7,484



<참여기업의 업종별 분류>



<참여기업수 상위 10개 프로그램>

사례 3: 기업체 교육프로그램 추천

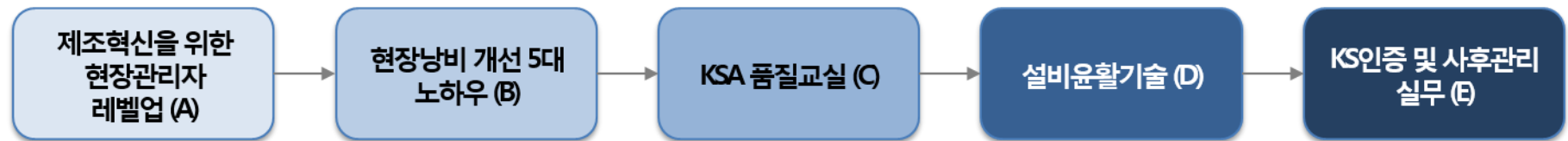
❖ 연관규칙분석 수행

- 지지도와 신뢰도의 기준값을 변화시켜 가며 총 25가지의 조합에 대해 생성된 규칙을 조사
- 현실적으로 활용 가능한 20~40개의 규칙을 생성한 세 가지 조합에 대한 추가 분석 실시
 - ✓ Case 1: 지지도 0.01, 신뢰도 0.2, Case 2: 지지도 0.01, 신뢰도 0.2, Case 3: 지지도 0.005, 신뢰도 0.3

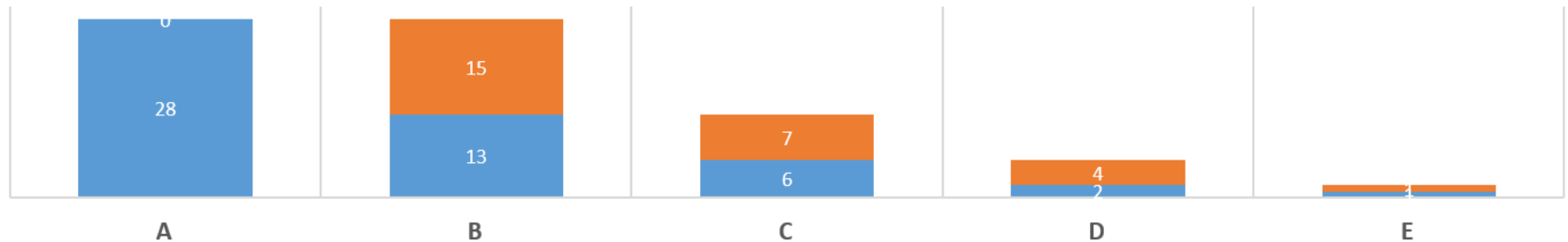
		지지도(support)				
		0.002	0.005	0.01	0.015	0.02
신뢰도(confidence)	0.1	2,264	196	40	19	6
	0.2	1,363	106	24	12	4
	0.3	928	40	4	2	1
	0.4	688	16	3	1	1
	0.5	592	13	2	1	1

사례 3: 기업체 교육프로그램 추천

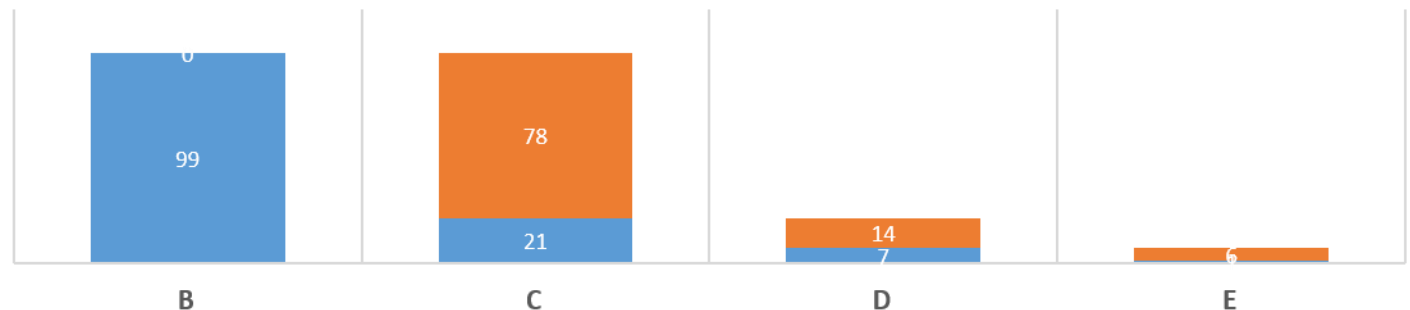
❖ 생성 규칙 분석 I: Chain 규칙 발굴을 통한 교육프로그램 단계적 추천



■ 수강 ■ 미수강



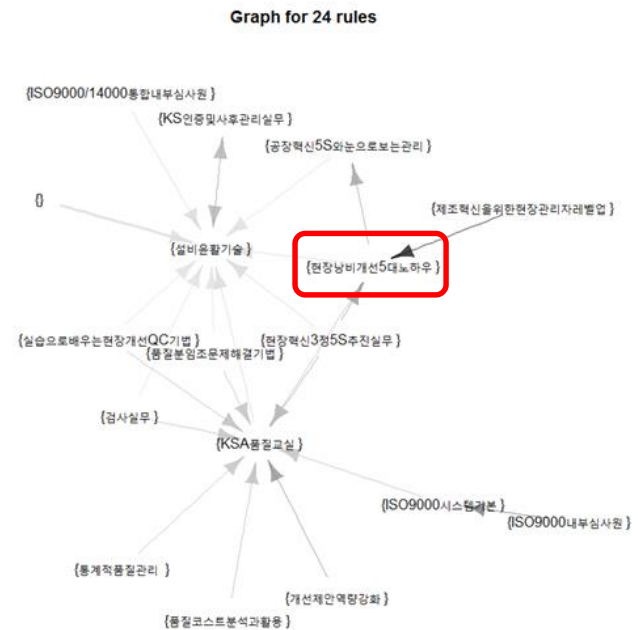
■ 수강 ■ 미수강



사례 3: 기업체 교육프로그램 추천

❖ 생성 규칙 분석 2: 수강기업수에 비해 규칙 등장 횟수가 높은 교육프로그램 발굴

순위	교육명	수강 기업수	순위	교육명	규칙등장 횟수
1	설비유효기술	1211	1	설비유효기술	11
2	KS인증및사후관리실무	234	2	KSA품질교실	10
3	KSA품질교실	201	3	현장낭비개선5대노하우	5
4	공장혁신5S와눈으로보는관리	148	4	현장혁신3정5S추진실무	3
5	ISO9000/14000통합내부심사원	122	5	품질분임조문제해결기법	2
6	ISO9000시스템기본	117	6	실습으로배우는현장개선QC기법	2
7	실습으로배우는현장개선QC기법	107	7	공장혁신5S와눈으로보는관리	2
8	품질분임조문제해결기법	103	8	검사실무	2
9	현장낭비개선5대노하우	99	9	KS인증및사후관리실무	2
10	검사실무	93	10	ISO9000시스템기본	2



사례 4: 해양 구조물 검사문서를 활용한 예방보수

❖ 국내 D사 사례

■ 배경

- ✓ 검사 시행 이후 조건부 합격의 부산물로 미흡한 부분에 대한 수정 사항
- ✓ 100여명의 검사원, 1인당 1일 300여 건의 검사문서 입력 (1년에 1천만 건)
- ✓ 한 개의 구조물에서 약 2년 간 약 28,800여 건이 입력



problem description	issue date	location stage	discipline priority
Sail Away 필수사항 PAGA - BV Cert (AO-040-HU-100-399010) PA/GA system is to be certified by BV. The comment will be cleared when BV certification is finalized	20101215	Hull	ELEC.
Sail Away 필수사항 PAGA, 75908025-EMC Test Report EMC test reports comments (LPO43521): Conducted emissions measurement, the limits specified in BV rules are exceeded at two distinct frequencies. To clarify if a passive filter can be adapted to suppress these frequencies / Transtel Vendor에서 관련 자료 BV Pass에 보냄 검토 승인대기 중부합조.	20101215	Hull	ELEC.
Freeboard Mark (AO-040-HU-100-202011) Load line mark assignment is pending to the receipt of the last revision of T & S Booklet Doc. No. AO-040-HU-100-202004	20101215	Hull	OUTF.
Ballast Water Management Plan (AO-040-HU-100-202020) Review is pending as follow: 1. Ballast exchange sequences are to be described precisely and relevant calculation concerned the effects of FSM & GM values to be submitted for review. 2. Loading condition used as initial step for the ballast water exchange sequences has been provided in the scope of intact stability with Trim & Stability Booklet	20101215	Hull	OUTF.
Strength Analysis for Mooring Equipment Foundations(Horizontal Shears) BV comment of stress criteria & correction of yield stress of 355 Mpa has mentioned in the calculation note "spread mooring equipment FONS on hull deck" will be closed after receipt of the final report.	20101215	Hull	OUTF.
		MECH-COMP.	B

사례 4: 해양 구조물 검사문서를 활용한 예방보수

❖ 국내 D사 사례

■ 목적

"S2 Upper Deck. No foundation structure member is installed for the panel structure, only one leg is welded."	
P6 Process Deck- Damaged installed pipeline (Fire water line Cu-Ni) Line No: 6045-FW-01405-647UP	
P1 (B129).The stair connection gusset plate is installed wrong position compare to company standard drawing	
"The installed stair handrail is different material, shall be changed to 25Φ pipe. (Handrail Post should be checked)"	
"P1 Module / The height of stair hand rail should be same (Standard drawing no: AO-040-TO-100-350042),The company sta...	
"P1 Module / The height of stair hand rail should be same (Standard drawing no: AO-040-TO-100-350042),The company sta...	
Beside of 148 block main stair landing below structure member mismatch to company standard drawing (Drawing no: AO-04...	
"Block 147 inside ceiling installed duct support mismatch uni-rack , Need to rearrange uni-rack (HVAC)"	
Front of duct	
Walkways to	
Stair legs to l	IZ-6001A/...
Junction Box	
Hydraulic stal	01A/B/C/...
Bolt missing for the mineral oil system. Equipment Tag No.: TO-UZ-6001A/B/C/D/E (8EA BOLT중 4EA씩 Missing)	
Monorails TO-JG3-6001 (S1) and TO-JG6-6001 (S2) to be upgraded to lift HSPT rotor and cartridge. Equipment Tag No.: TO-...	
Electric shortage on Chilled water unit - Cable and panel burn damaged.Equipment No: TO-UE-0921A/B CHILLER	
Pipe support is installed on Module lifting lug. Lift up interference with lifting device during module lift up. The pipe support sh...	
P3 Module : Padeye location for Butterfly valve handling	
S2 Module - Too low at Instrument Air manifold drain point. 6045-2-IA-TO-07434-B49-272TP-P01	
"Need to modify these lines due to so complex at marking area in DO lines of the Turbo Generator ""E"" in the Process Deck."	
Support of GRP Piping to be installed underneath of the process deck	
"On the S1 & S2 Mezzanine deck, the pipe support does not match with structure and grating"	
Electrical Support - Type of support to be changed	
Drain holes to be provided at middle part duct flange. of inlet duct (FN33) for WIP.	

From Textual Documents To Insight

언제 어디서 어떠한 유형의 문제가 발생하는가?

■ 분석 결과

replac		cover	
change	cable	valve	
tube	need	point	ventil
panel	pipe	bolt	line
		connect	

“손상이 발생하는 부품들은?”

40/62

사례 4: 해양 구조물 검사문서를 활용한 예방보수

❖ 국내 D사 사례

■ 분석 결과

“pipe”와 함께 등장하는 단어

pipe	support	line	page	cable	guide	addition	ISO	grate	leak	equipment	welding	deformation	inspection	padding	replacement	action
		install	add	shake	deck	adjust	pad	connection	location	intrusion	requirement	identification	film	resin	fit	test
	gap	weld	joint	tray	keep	DWG	end	connection	bar	bolt	chamber	up	of	middle	to	ad
		point	modify	confirm	clash	miss	shoe	structure	touch	alignment	fix	tubing	loose	plastic	cl	

“pipe”, “support”와 함께 등장하는 단어

pipe	add	page	point	DWG	correct	deck	pad	spool	confirm	shoe
				cable	guide	miss	end	fix	middle	cradle
support	gap	install	addition	clash	cha	loc	ne	pla	su	val
				in...	hull	rei	ali	co	an	int
	shake	weld	joint	ISO	bolt	sm	n	a	b	
				beam	c...	mo	to	p	b	

사례 4: 해양 구조물 검사문서를 활용한 예방보수

❖ 국내 D사 사례

■ 분석 결과

“pipe”, “support”와 함께 등장하는 단어

pipe	add	page	point	DWVG	correct	deck	pad	spool	confi	shoe
				cabl	guid	miss	end	fix	mid	crad
support	gap	install	addition	cl...	cha	loc	ne	pla	su	val
				clash	rei	ali	co	an	int	lig
	shake	weld	joint	ISO	hull	fla	of	s	st	v
				bolt	sm	pl	n	a		
		line	modify	bea	mo	to	p	b		
				m	c...	ce	r	d		
		adjust		type	le...	re	cl	r	g	
				tray	s...	pro	co	s	h	
				teflo	in...	dra	cu	s	o	
				n	k...	GR	dr	u	s	

notc_detl_desc

(8506-31S-TW-905) Add pipe support in the middle of [1-2] spool due to...

(2902-51S-VE-003) Additional support to be installd in order no to shak...

(0 (2902-51S-VE-003) Additional support to be installd in order no to shake of the upper part of the pipe spool. (1"-VE-TO-29251-B49-577TP)

CFR-TO-0350 / S3# Top deck The additional support to be installed du...

Add a additional pipe support at the middle of spool to avoid shaking.

Additional pipe support is required at the middle of spool to avoid shaki...

Add an additional pipe support at the end of pipe spool [1-4] Valve side ...

Add additional pipe support at all 2" Elbow sides due to shaking (10 pla...

Add additional pipe support near weld joint-39 on page-006 and weld joi...

“pipe”, “support”, “add”가 함께 등장하는 검사문서

“shake”도 함께 등장함

→ “파이프(pipe)가 흔들리니(shake),
지주(support)를 추가(add)하라”

사례 4: 해양 구조물 검사문서를 활용한 예방보수

❖ 국내 D사 사례

■ 분석 결과

“파이프(pipe)가 흔들리니(shake), 지주(support)를 추가(add)하라”

언제, 어디서?

“공사 중반, Topside General, Hull, P5,6,7 Module 등에서 주로 발생한다”

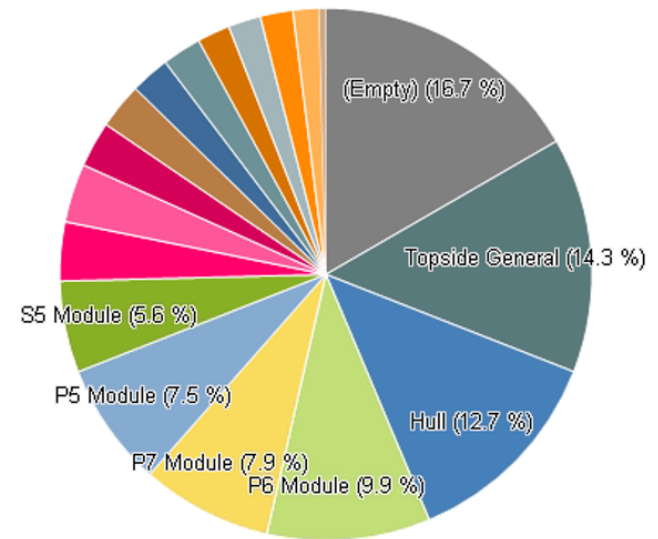
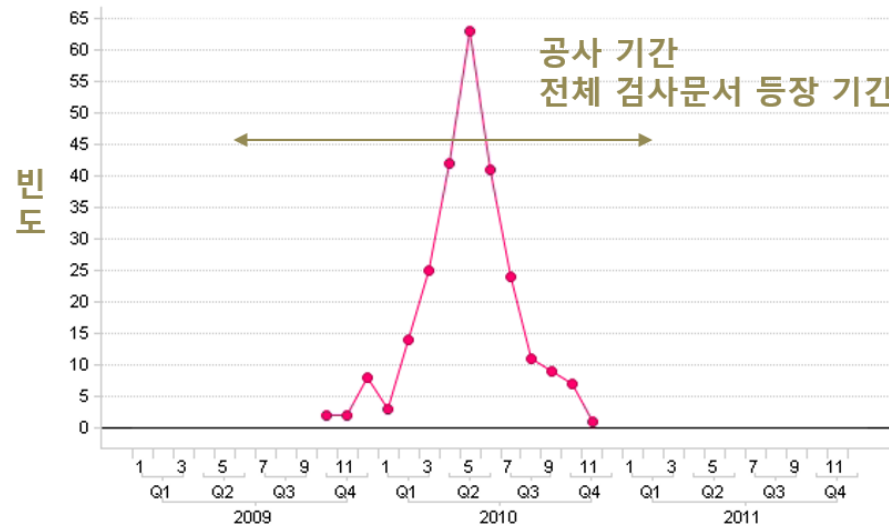


Table of Contents

I

Association Rules:Theory

II

Association Rules:Applications

III

R Exercise

Association Rules

❖ Package “arules” & “arulesViz”

Package ‘arules’

February 21, 2014

Version 1.1-2

Date 2014-2-21

Title Mining Association Rules and Frequent Itemsets

Description Provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Also provides interfaces to C implementations of the association mining algorithms Apriori and Eclat by C. Borgelt.

Classification/ACM G.4, H.2.8, I.5.1

URL <http://R-Forge.R-project.org/projects/arules/>,
<http://lyle.smu.edu/IDA/arules/>

Depends R (>= 2.14.2), Matrix (>= 1.0-0)

Imports stats, methods

Suggests pmml, arulesViz, testthat

Package ‘arulesViz’

March 11, 2014

Version 0.1-9

Date 2014-03-10

Title Visualizing Association Rules and Frequent Itemsets

Author Michael Hahsler and Sudheer Chelluboina

Maintainer Michael Hahsler <mhahsler@lyle.smu.edu>

Depends R (>= 2.14.0), arules (>= 1.0-5), grid

Imports scatterplot3d, vcd, seriation, igraph

Suggests iplots, Rgraphviz, graph

Description Various visualization techniques for association rules and itemsets. The packages also includes several interactive visualizations for rule exploration. This package extends package arules.

Association Rules: Data Frame

❖ Titanic Data

```

1 # Association Rules -----
2 # arules and arulesviz packages install
3 install.packages("arules", dependencies = TRUE)
4 install.packages("arulesviz", dependencies = TRUE)
5
6 library(arules)
7 library(arulesviz)
8 library(wordcloud)
9
10 # Load titanic data set
11 titanic <- read.delim("titanic.txt", dec=",")
12 str(titanic)
13 head(titanic)

> # Load titanic data set
> titanic <- read.delim("titanic.txt", dec=",")
> str(titanic)
'data.frame': 1313 obs. of  5 variables:
 $ Name      : Factor w/ 1310 levels "Abbing, Mr Anthony",...: 22 25 26 27 24 31 45 46 50 54 ...
 $ Pclass    : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
 $ Age       : Factor w/ 75 levels "0.17","0.33",...: 28 18 30 24 5 48 66 39 60 73 ...
 $ Sex       : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
 $ Survived: int  1 0 0 0 1 1 1 0 1 0 ...
> head(titanic)

```

	Name	Pclass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29	female	1
2	Allison, Miss Helen Loraine	1st	2	female	0
3	Allison, Mr Hudson Joshua Creighton	1st	30	male	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25	female	0
5	Allison, Master Hudson Trevor	1st	0.92	male	1
6	Anderson, Mr Harry	1st	47	male	1

Association Rules: Data Frame

❖ Data Preprocessing

- Categorize a numeric variable, remove NA, etc.

```

15 # Remove "Name" column and group "Age" column
16 titanic_ar <- titanic[,2:5]
17 titanic_ar$Age = as.character(titanic_ar$Age)
18 c_idx <- which(as.numeric(titanic_ar$Age) < 20)
19 a_idx <- which(as.numeric(titanic_ar$Age) >= 20)
20 na_idx <- which(is.na(titanic_ar$Age))
21
22 titanic_ar$Age[c_idx] <- "Child"
23 titanic_ar$Age[a_idx] <- "Adult"
24 titanic_ar$Age[na_idx] <- "Unknown"
25
26 # Convert the attributes to factor
27 titanic_ar$Age <- as.factor(titanic_ar$Age)
28 titanic_ar$Survived <- as.factor(titanic_ar$Survived)

```

	PClass	Age	Sex	Survived
1	1st	Adult	female	1
2	1st	Child	female	0
3	1st	Adult	male	0
4	1st	Adult	female	0
5	1st	Child	male	1
6	1st	Adult	male	1
7	1st	Adult	female	1
8	1st	Adult	male	0
9	1st	Adult	female	1
10	1st	Adult	male	0

Association Rules: Data Frame

❖ Find rules (default setting)

```
30 # Rule generation by Apriori algorithm with default settings
31 rules <- apriori(titanic_ar)
32 inspect(rules)
```

```
> rules <- apriori(titanic_ar)
```

parameter specification:

confidence	minval	smax	arem	aval	originalsupport	support	minlen	maxlen	target	ext
0.8	0.1	1	none	FALSE	TRUE	0.1	1	10	rules	FALSE

algorithmic control:

filter	tree	heap	memopt	load	sort	verbose
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

```
apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)      (c) 1996-2004  Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10 item(s), 1313 transaction(s)] done [0.00s].
sorting and recoding items ... [10 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [16 rule(s)] done [0.00s].
creating s4 object ... done [0.00s].
```

Association Rules: Data Frame

❖ Find rules (default setting)

```
> inspect(rules)
```

	lhs	rhs	support	confidence	lift
1	{PClass=3rd}	=> {Survived=0}	0.4364052	0.8059072	1.226137
2	{Sex=male}	=> {Survived=0}	0.5399848	0.8331375	1.267566
3	{Survived=0}	=> {Sex=male}	0.5399848	0.8215527	1.267566
4	{PClass=2nd, Sex=male}	=> {Survived=0}	0.1127190	0.8554913	1.301576
5	{PClass=2nd, Survived=0}	=> {Sex=male}	0.1127190	0.9192547	1.418309
6	{PClass=1st, Sex=female}	=> {Survived=1}	0.1020564	0.9370629	2.734141
7	{Sex=female, Survived=0}	=> {PClass=3rd}	0.1005331	0.8571429	1.582881
8	{PClass=3rd, Age=Unknown}	=> {Survived=0}	0.2536177	0.8473282	1.289156
9	{Age=Unknown, Sex=male}	=> {Survived=0}	0.2566641	0.8798956	1.338706
10	{Age=Unknown, Survived=0}	=> {Sex=male}	0.2566641	0.8023810	1.237986
11	{Age=Adult, Sex=male}	=> {Survived=0}	0.2482864	0.8253165	1.255667
12	{Age=Adult, Survived=0}	=> {Sex=male}	0.2482864	0.8693333	1.341286
13	{PClass=3rd, Sex=male}	=> {Survived=0}	0.3358720	0.8837675	1.344596
14	{PClass=3rd, Age=Unknown, Sex=male}	=> {Survived=0}	0.1957350	0.9081272	1.381658
15	{PClass=3rd, Age=Adult, Sex=male}	=> {Survived=0}	0.1142422	0.8670520	1.319165
16	{PClass=3rd, Age=Adult, Survived=0}	=> {Sex=male}	0.1142422	0.8064516	1.244267

Association Rules: Data Frame

❖ Find rules (customized setting)

```

34 # Rule generation by Apriori algorithm with custom settings
35 rules <- apriori(titanic_ar, parameter = list(minlen = 3, support = 0.1, conf = 0.8),
36               appearance = list(rhs = c("Survived=0", "Survived=1"), default="lhs"))
37 inspect(rules)
38
39 # Plot the rules
40 plot(rules, method="scatterplot")
41 plot(rules, method="graph", control=list(type = "items", alpha = 1))
42 plot(rules, method="paracoord", control=list(reorder=TRUE))

```

```

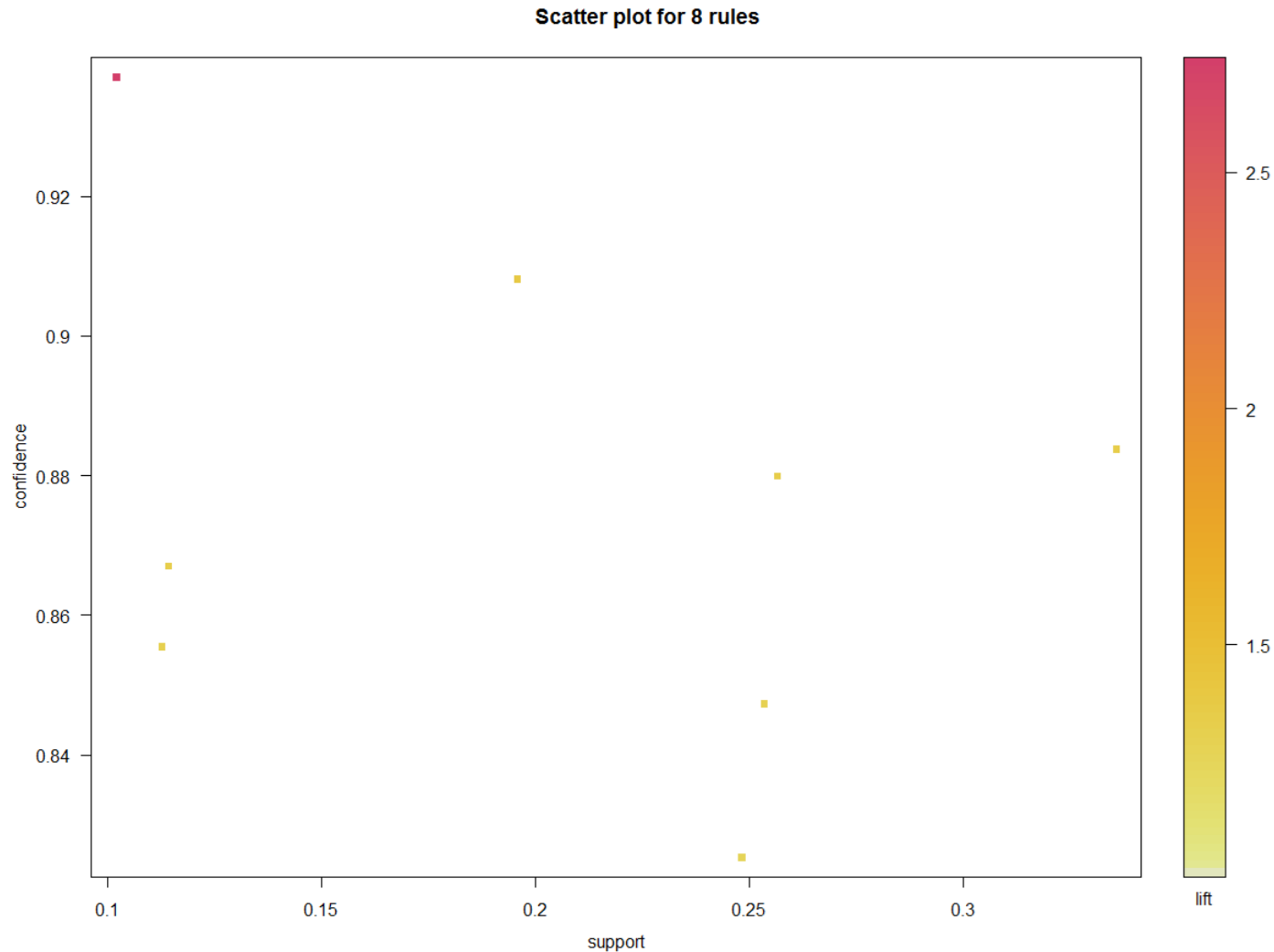
> inspect(rules)

```

	lhs	rhs	support	confidence	lift
1	{Pclass=2nd, Sex=male}	=> {Survived=0}	0.1127190	0.8554913	1.301576
2	{Pclass=1st, Sex=female}	=> {Survived=1}	0.1020564	0.9370629	2.734141
3	{Pclass=3rd, Age=Unknown}	=> {Survived=0}	0.2536177	0.8473282	1.289156
4	{Age=Unknown, Sex=male}	=> {Survived=0}	0.2566641	0.8798956	1.338706
5	{Age=Adult, Sex=male}	=> {Survived=0}	0.2482864	0.8253165	1.255667
6	{Pclass=3rd, Sex=male}	=> {Survived=0}	0.3358720	0.8837675	1.344596
7	{Pclass=3rd, Age=Unknown, Sex=male}	=> {Survived=0}	0.1957350	0.9081272	1.381658
8	{Pclass=3rd, Age=Adult, Sex=male}	=> {Survived=0}	0.1142422	0.8670520	1.319165

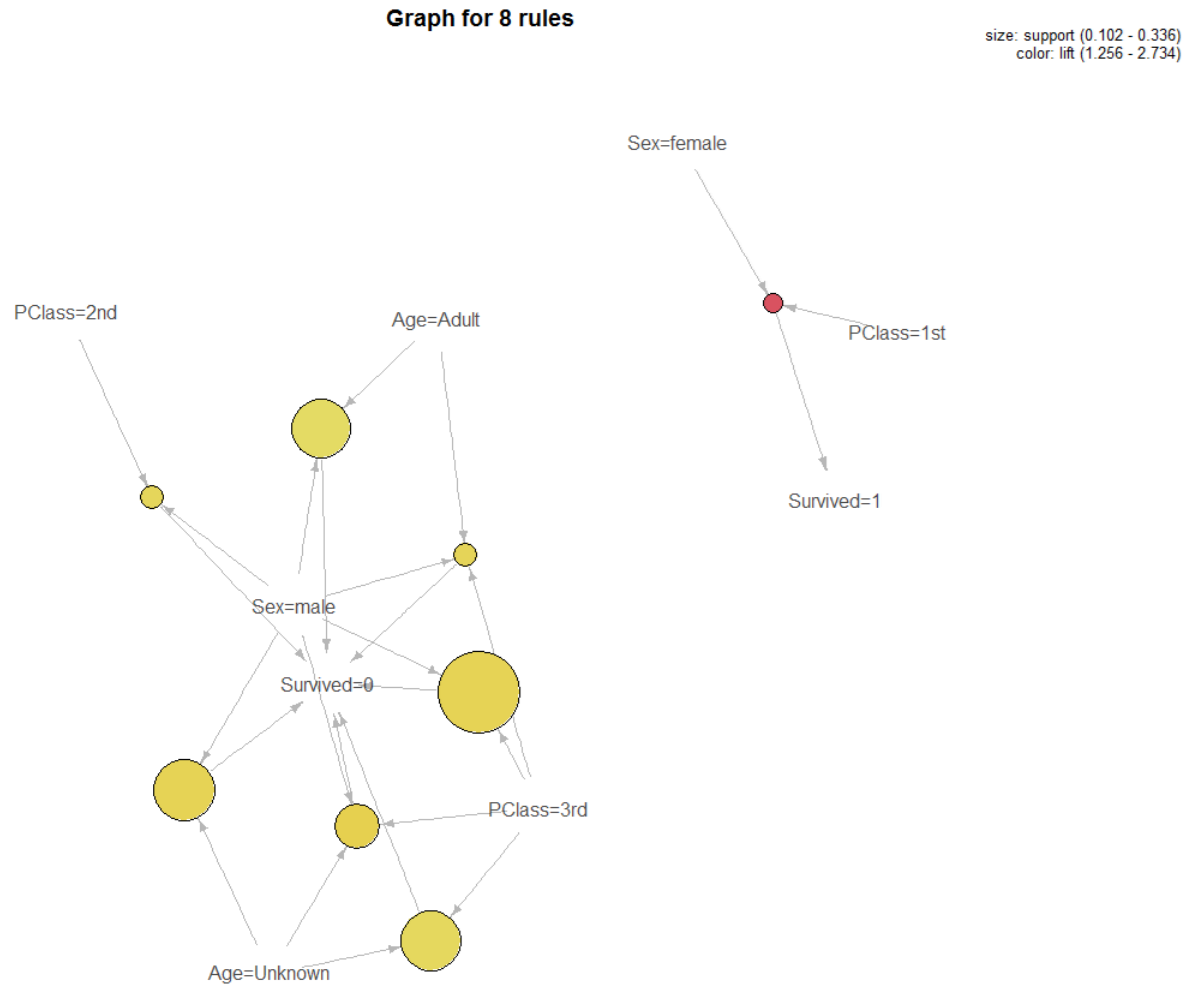
Association Rules: Data Frame

❖ Visualize the rules



Association Rules: Data Frame

❖ Visualize the rules



Parallel coordinates plot for 8 rules



Association Rules: Transaction Data

❖ Groceries shopping data

```
44 # Load transaction data "Groceries"
45 data("Groceries")
46 summary(Groceries)
47 str(Groceries)
48 inspect(Groceries)
```

```
> summary(Groceries)
```

```
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146
```

```
most frequent items:
```

whole milk	other vegetables	rolls/buns	soda	yogurt	(other)
2513	1903	1809	1715	1372	34055

```
element (itemset/transaction) length distribution:
sizes
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2159	1643	1299	1005	855	645	545	438	350	246	182	117	78	77	55	46	29	14	14	9	11	4	6	1
26	27	28	29	32																			
1	1	1	3	1																			

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	4.409	6.000	32.000

```
includes extended item information - examples:
```

	labels	level2	level1
1	frankfurter	sausage	meet and sausage
2	sausage	sausage	meet and sausage
3	liver loaf	sausage	meet and sausage

Association Rules: Transaction Data

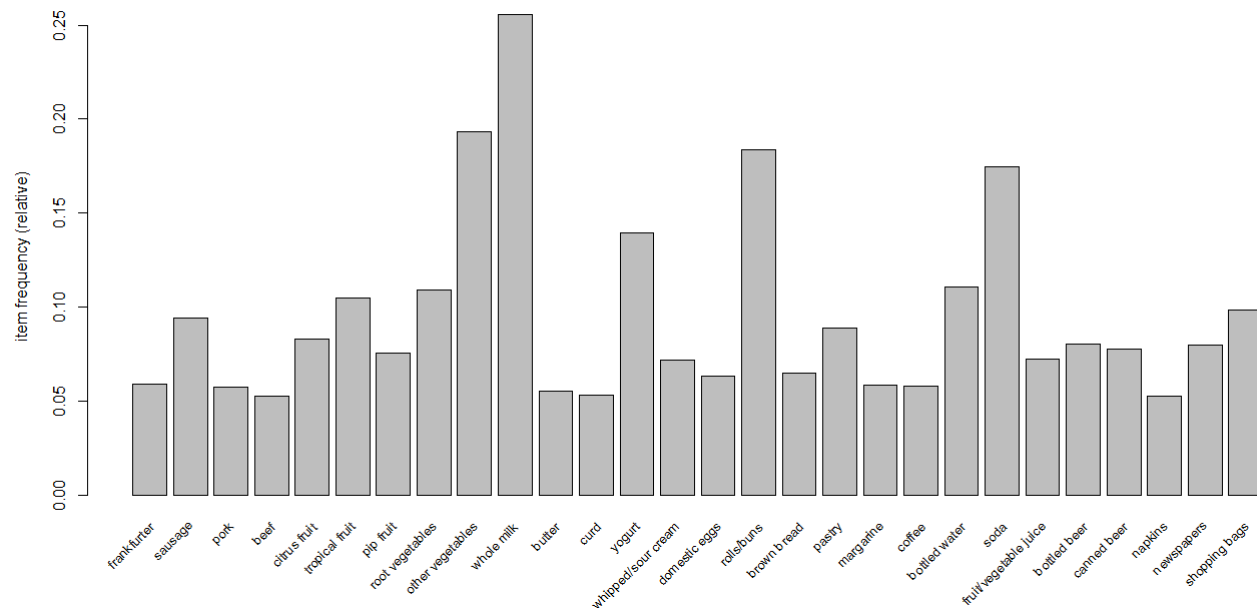
❖ Groceries shopping data

■ Item inspection

```

50 # Item inspection
51 itemName <- itemLabels(Groceries)
52 itemCount <- itemFrequency(Groceries)*9835
53
54 col <- brewer.pal(8, "Dark2")
55 wordcloud(words = itemName, freq = itemCount, min.freq = 1, scale = c(10, 0.2), col = col , random.order = FALSE)
56
57 itemFrequencyPlot(Groceries, support = 0.05, cex.names=0.8)

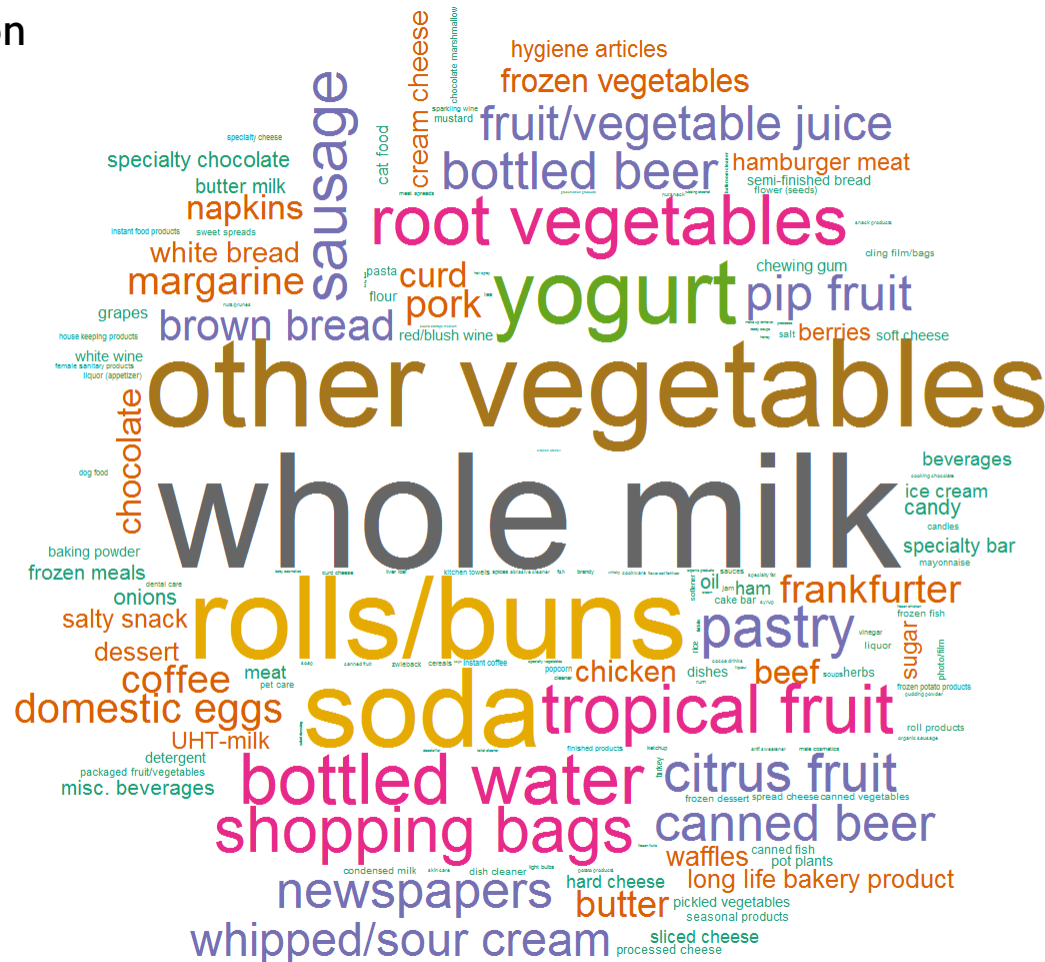
```



Association Rules: Transaction Data

❖ Groceries shopping data

- Item inspection



Association Rules: Transaction Data

❖ Find and visualize rules

```

59 # Rule generation by Apriori
60 rules <- apriori(Groceries, parameter=list(support=0.001, confidence=0.5))
61 rules
62
63 # List the first three rules with the highest lift values
64 inspect(head(sort(rules, by="lift"),3))
65
66 # Save the rules in a text file
67 write.csv(as(rules, "data.frame"), "Groceries_rules.csv", row.names = FALSE)
68
69 # Plot the rules
70 plot(rules)
71 plot(rules, method="grouped")

```

```

> inspect(head(sort(rules, by="lift"),3))

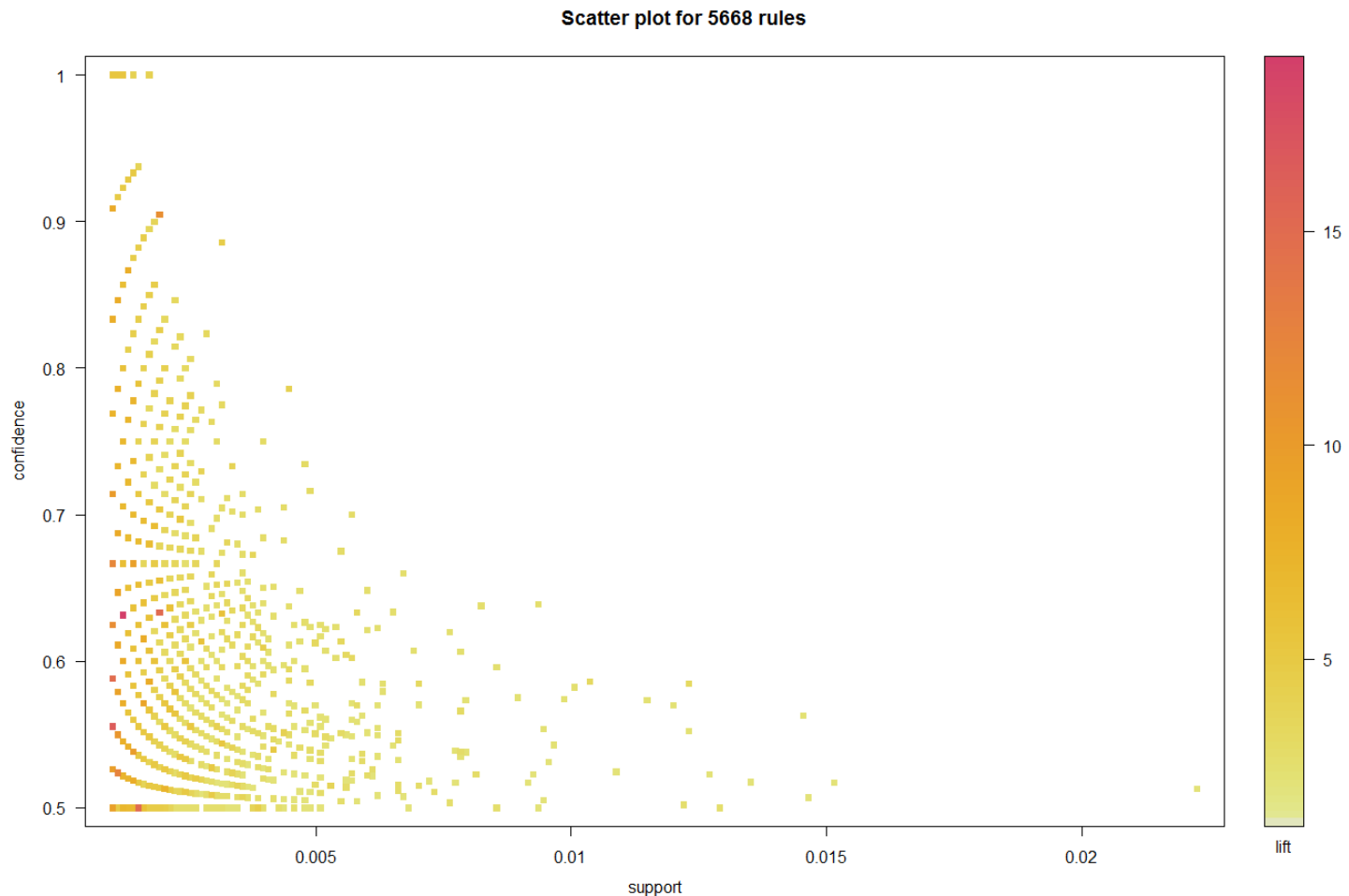
```

	lhs	rhs	support	confidence	lift
1	{Instant food products, soda}	=> {hamburger meat}	0.001220132	0.6315789	18.99565
2	{soda, popcorn}	=> {salty snack}	0.001220132	0.6315789	16.69779
3	{flour, baking powder}	=> {sugar}	0.001016777	0.5555556	16.40807

	A	B	C	D
1	rules	support	confidence	lift
2	{honey} => {whole milk}	0.001118	0.733333	2.870009
3	{tidbits} => {rolls/buns}	0.00122	0.521739	2.836542
4	{cocoa drinks} => {whole milk}	0.001322	0.590909	2.312611
5	{pudding powder} => {whole milk}	0.001322	0.565217	2.212062
6	{cooking chocolate} => {whole milk}	0.001322	0.52	2.035097
7	{cereals} => {whole milk}	0.00366	0.642857	2.515917
8	{jam} => {whole milk}	0.002949	0.54717	2.141431
9	{specialty cheese} => {other vegetables}	0.00427	0.5	2.584078
10	{rice} => {other vegetables}	0.003965	0.52	2.687441
11	{rice} => {whole milk}	0.004677	0.613333	2.400371
12	{baking powder} => {whole milk}	0.009253	0.522989	2.046793
13	{liver loaf,yogurt} => {whole milk}	0.001017	0.666667	2.609099
14	{tropical fruit,curd cheese} => {other vegetables}	0.001017	0.666667	3.445437
15	{curd cheese,rolls/buns} => {whole milk}	0.001017	0.625	2.446031
16	{other vegetables,curd cheese} => {whole milk}	0.00122	0.571429	2.236371
17	{whole milk,curd cheese} => {other vegetables}	0.00122	0.521739	2.696429
18	{other vegetables,cleaner} => {whole milk}	0.001017	0.625	2.446031
19	{liquor,red/blush wine} => {bottled beer}	0.001932	0.904762	11.23527
20	{soda,liquor} => {bottled beer}	0.00122	0.571429	7.09596

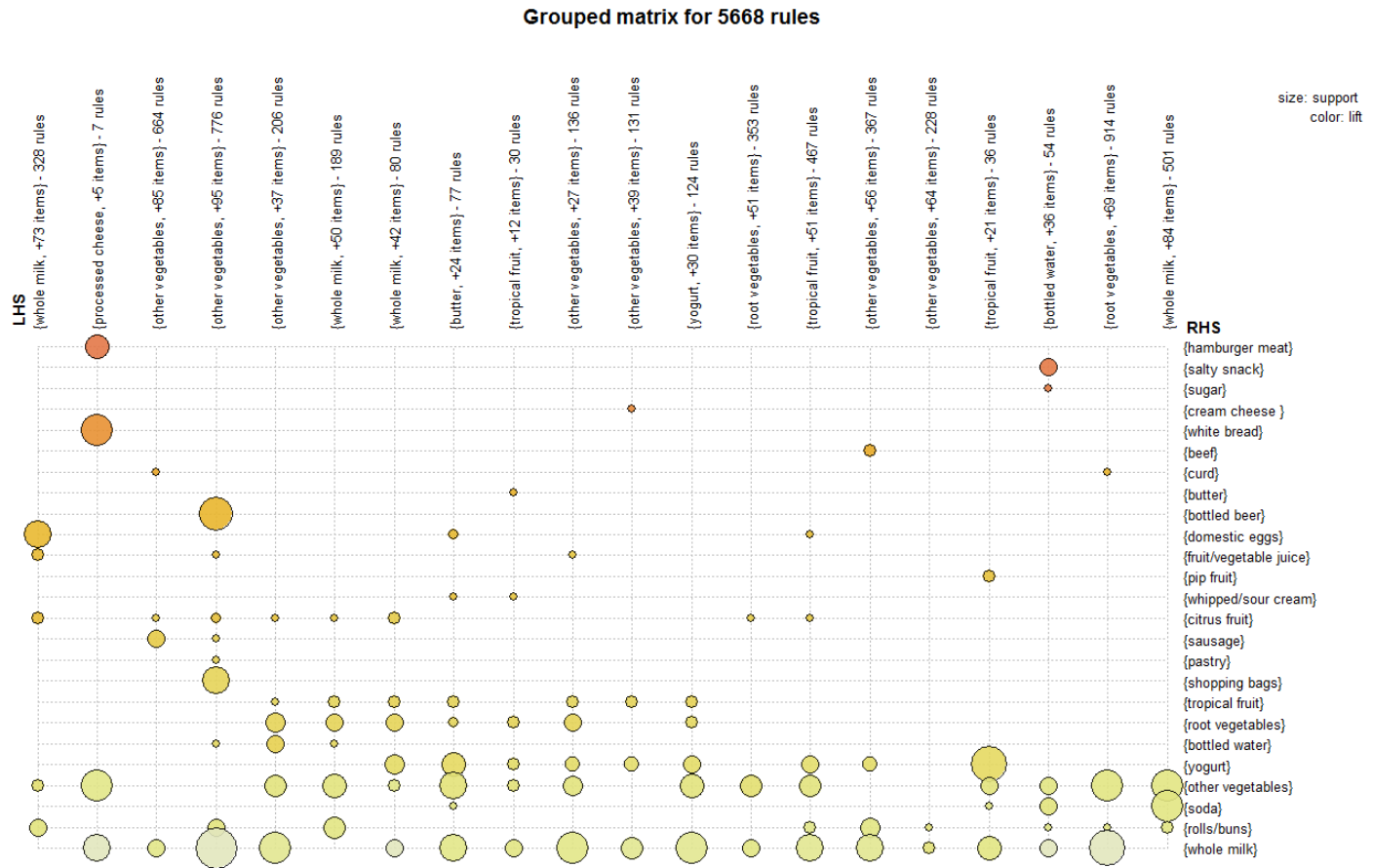
Association Rules: Transaction Data

❖ Find and visualize rules



Association Rules: Transaction Data

❖ Find and visualize rules



Q & A

